

DEVELOPMENT OF DATA QUALITY CONTROL LIMITS FOR DATA SCREENING THROUGH THE “ENERGY BALANCE” METHOD

Hiroko Masuda
Graduate Research Assistant

Juan Carlos Baltazar, Ph.D.
Associate Research Engineer

Jing Ji, Ph.D.
Research Associate

David E. Claridge, Ph.D., P.E.
Professor / Director

Energy Systems Laboratory, Texas Engineering Experiment Station
Texas A&M University, College Station, TX

ABSTRACT

Energy management processes such as accounting energy cost, finding overuse in energy, and determining savings from energy conservation programs largely depends on measured energy use data. Identifying and correcting faulty data properly would avoid over/underestimation in energy use and increase accuracy in further analysis. It allows engineers and administrators to make more confident and low-risk decisions.

This paper proposes a methodology to construct statistical control limits for data screening using “Energy Balance” methodology (Shao and Claridge, 2005). Energy Balance (E_{BL}) parameter represents quasi-steady state thermal energy storage in a building and indicates a predominant linear behavior when it is plotted versus the outside air temperature. A regression model of E_{BL} parameter developed as a function of the outside air temperature from a long-term data can be used as a data screening tool for newly measured energy use in the building. However, E_{BL} model is known to have functional discontinuities called “change points” and non-uniform residuals. To construct control limits that fit the E_{BL} data uniformly over a wide range of outside air temperature, a new technique was introduced which estimate mean square error (MSE) for a change point model as a function of outside air temperature by using *Bin-MSE* data.

This methodology has successfully constructed data quality control limits of E_{BL} parameter for example buildings. The numerical criteria developed by this methodology would help to have uniform results in quality control for energy use data, and it would be used as a rule for an automated data screening process.

INTRODUCTION

Energy Balance methodology (Shao and Claridge, 2005) is an innovative data screening technique based on the first law of thermodynamics

in conjunction with the concept of analytical redundancy. This methodology defines Energy Balance (E_{BL}) parameter which represents quasi-steady state thermal energy storage in a building. It was examined that E_{BL} parameter is dependent of the outside dry-bulb temperature and follows a predominant linear behavior. By knowing the pattern of the E_{BL} parameter for a building evaluated by the measured electricity, chilled water and heating hot water use during reference period as a function of the outside air temperature, it is possible to verify the newly observed energy use data in the building under the same operating condition as in the reference period.

This technique has been applied to approximately 100 buildings on the Texas A&M University main campus for data quality control of energy use data, and has proved to be an effective data quality screening method for verification of sensors for whole building energy use (Baltazar et al., 2007). In the quality control process, E_{BL} parameter is evaluated and plotted as a function of the outside air temperature (E_{BL} plot). Then the data analysis specialists refer the E_{BL} plot with the corresponding time series plot of each of the three series of energy use data: electricity, chilled water, and heating hot water to detect pattern troubles, misbehavior or unusual data visually. This visual analysis has detected faulty data such as errors in scale of the recorded data, errors due to the apparent malfunction of the sensors very well. However, visual detection depends on specialists’ experiences and causes variations in the result and reliability. Therefore, scientific and numerical criteria of data quality control are required in addition to the visual analysis. These criteria can be used as rules for an automated data screening tool in the future, and it would reduce analysis time when many buildings have to be analyzed at one time.

This paper presents a methodology to set quality control limits for E_{BL} plot so that the plot with the

control limits can be used as similar to well-known “process control chart”. The control limits for E_{BL} parameter are defined by extending the concept of prediction error in linear regression models to change point models with non-uniform residuals. This methodology provides control limits on E_{BL} plots under prescribed uncertainty level for the purpose of data screening.

ENERGY BALANCE PARAMETER GENERAL FORMULATION

The general derivation of the “Energy Balance” parameter comes from the first law of thermodynamics. The process is modeled as a semi-empirical methodology based on analytic redundancy (Shao and Claridge, 2006). For a whole building thermodynamic model, the rates of heat flow and the rates of enthalpy flow across the boundary of the control volume and the rates of work performed on the control volume may be broken into seven major components: internal heat gain from lighting and equipment (fWb_{ele}), heating provided to the building by the HVAC system (Wb_{heat}), heat removed from the building by the cooling system (Wb_{cool}), solar radiation through the envelope (Q_{solar}), ventilation air and infiltration via doors, windows, or air-handling units, (Q_{vent}), heat transmission through the building structure (Q_{cond}), and heat gain from occupants (Q_{occ}). The energy balance equation for a building becomes:

$$\frac{d}{dt} \dot{E} = \dot{Q}_{vent} + \dot{Q}_{solar} + \dot{Q}_{cond} + \dot{Q}_{occ} + \dot{W}_{bheat} - \dot{W}_{bcool} + f\dot{W}_{bele} \quad (1)$$

where E is the thermal energy storage in the building. The term fWb_{ele} is the fraction of whole building electricity that is non-cooling use and converted to internal gain. fWb_{ele} , Wb_{cool} , and Wb_{heat} are separately metered and monitored in the buildings. If the analysis is made on the basis of a period greater than a day, Eq. 1 can be considered quasi-steady. Then Energy Balance parameter, E_{BL} is defined as the relationship between metered terms:

$$E_{BL} = \dot{W}_{bheat} - \dot{W}_{bcool} + f\dot{W}_{bele} = -(\dot{Q}_{vent} + \dot{Q}_{solar} + \dot{Q}_{cond} + \dot{Q}_{occ}) \quad (2)$$

Shao and Claridge (2006) have examined characteristic and sensitivity of E_{BL} parameter by simulation and provided that the E_{BL} parameter is independent of the type of air handling unit that is used in the building HVAC system, and operational conditions such as cooling coil leaving air temperature and outside air intake volume are the

key parameters that strongly influence the simulated values of E_{BL} .

QUALITY CONTROL METHODOLOGY USING CONTROL CHART

The purpose of Statistical Quality Control (QC) is to detect and reduce variability in the process. ‘Control Chart’ is effective method to determine variability and detect problematic points. Figure 1 shows a typical control chart for a single quality characteristic. The quality characteristics are measured or computed from a sample and plotted on the chart versus the sample number or versus time. In general, the chart consists of a center line (CL), an upper control limit (UCL) and a lower control limit (LCL). CL represents the average value of the quality characteristic corresponding to the in-control state. UCL and LCL are chosen so that if the process is in control, nearly all of the sample points will fall between them. In general, as long as the points plot within the control limits, the process is assumed to be in control, and no action is necessary. However, a point that plots outside of the control limits is interpreted as evidence that the process is out of control, and investigation and corrective action are required to find and eliminate the assignable cause of this behavior (Montgomery and Runger, 2003).

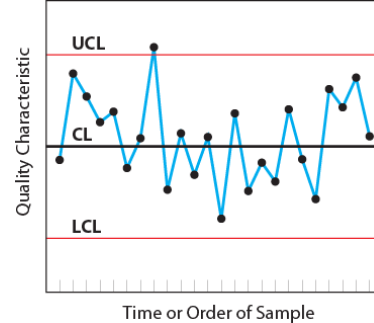


Figure 1 A typical control chart

Under the assumption that the individual data values are distributed normally and independently about the mean, a general estimation of control chart parameters is given in Eq. 3. The coverage factor k is a multiplier of the standard deviation that defines the control limits and can take any real number. The coverage factor implies the confidence level.

$$UCL = \bar{X} + kS_x \quad (3a)$$

$$CL = \bar{X} \quad (3b)$$

$$LCL = \bar{X} - kS_x \quad (3c)$$

where

\bar{X} : mean value of the sample

S_x : standard deviation of the sample
 k : coverage factor

For the quality control of energy use data using Energy Balance methodology, the definition of control chart is applied to E_{BL} parameter as a function of outside air temperature, and control chart variables are defined as Eq. 4 by assuming individual observations of E_{BL} are distributed normally and independently about the mean.

$$UCL(T) = \hat{E}_{BL}(T) + kS_{\hat{E}_{BL}} \quad (4a)$$

$$CL(T) = \hat{E}_{BL}(T) \quad (4b)$$

$$LCL(T) = \hat{E}_{BL}(T) - kS_{\hat{E}_{BL}} \quad (4c)$$

where

T : outside air temperature

$\hat{E}_{BL}(T)$: estimated mean response of E_{BL} at the outside air temperature T

$S_{\hat{E}_{BL}}$: estimated standard deviation of E_{BL}

The parameters for $E_{BL}(T)$ and $S_{\hat{E}_{BL}}$ can be calculated by E_{BL} evaluated by a sample of energy use data and outside air temperature from the reference period. Methodologies to estimate $\hat{E}_{BL}(T)$ and $S_{\hat{E}_{BL}}$ will be explained in the following sections.

Figure 2 shows a general procedure for data quality control using a control chart for E_{BL} defined in Eq. 4. The inputs are daily energy use data for three types of energy sources: electricity, chilled water, and heating hot water and the corresponding outside air temperature (T_{OA}). Missing data may be already filled by appropriate interpolation methods. In the next step, E_{BL} parameter is evaluated by the energy use data and plotted versus outside air temperature (E_{BL} plot) on the control chart for the building. Then the data is checked if it falls inside or outside the control limits. If a newly observed data point is within the bound of control limits, the data is classified to be acceptable, and if it is out of the bound, the data is assumed to have misbehavior or unusual pattern, and the cause should be identified. If it is needed to correct the data, it will be estimated by an appropriate method. After this quality control process, the data is qualified for further procedures such as determination of the consumption during a particular period.

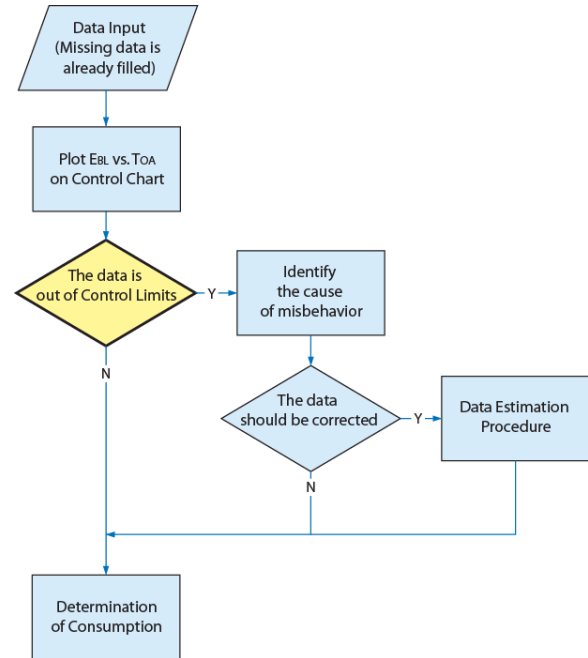


Figure 2 Block diagram of the quality control process

REGRESSION MODEL OF ENERGY BALANCE PARAMETER

The model parameters for mean response of E_{BL} can be estimated by regression analysis using the data during a reference period, by selecting E_{BL} as dependent and outside air temperature as independent variables.

Based on the analysis of the functionality of energy systems following physical principles, Shao and Claridge (2006) have shown the sensible portion of the Energy Balance parameter depends linearly on the outside air temperature and the average values of latent portion versus outside air temperature can be fit by a polynomial expression of order four. In total, E_{BL} shows predominant line behavior with a functional discontinuity which is called “change point”. At a change point, the line turns from linear into polynomial line.

Classic simple linear regression (SLR) is not appropriate for E_{BL} because of the change point, but this is not a new problem. It is known that energy use for buildings with continuous, year-round cooling or heating often has a change point due to the presence of control mechanisms. To achieve a better fitting in those cases, change point models have been introduced and studied well. Ruch and Claridge (1992) presented the application of four-parameter change-point (4P-CP) models as a function of dry-bulb temperature to building energy use, and uncertainty analysis for change point regression models has been done by Reddy, et al (1997 and

1998). The 4P-CP model algorithm has been adopted by several energy analysis tools as the EModel (Kissock et al., 1993, 2007) and the ASHRAE Inverse Modeling Toolkit (IMT) which was designed to model a wide variety of energy use patterns (Kissock et al. 2002). Generally, 4P-CP models show better fitting to E_{BL} than SLR models. The 4P-CP model of E_{BL} as a function of outside air temperature can be written as:

$$E_{BL} = E_{BL,CP} - LS(T_{CP} - T)^+ + RS(T - T_{CP})^+ \quad (5)$$

where

T : outside air temperature

T_{CP} : change point outside air temperature

LS : slope below the T_{CP}

RS : slope above the T_{CP}

$E_{BL,CP}$: E_{BL} associated with the T_{CP}

and the superscript plus sign indicates that only positive values are counted.

In this paper, 4P-CP model was estimated by *CPReg* (Baltazar 1999). This program calculates statistical and model parameters for Mean, SLR, 3P and 4P change point regressions, and provides the scatter plot of the data with the corresponding model. Figure 3 is a comparison of SLR and 4P-CP E_{BL} models. E_{BL} was evaluated from the actual data in an office building, and the models and data points were estimated and plotted by *CPReg*. 4P-CP model shows better fitting to the actual E_{BL} data over a wide range of outside air temperature. In this example, RMSE of SLR model is 65.5 while RMSE of 4P-CP model is 54.6 (Units Btu/day ft²).

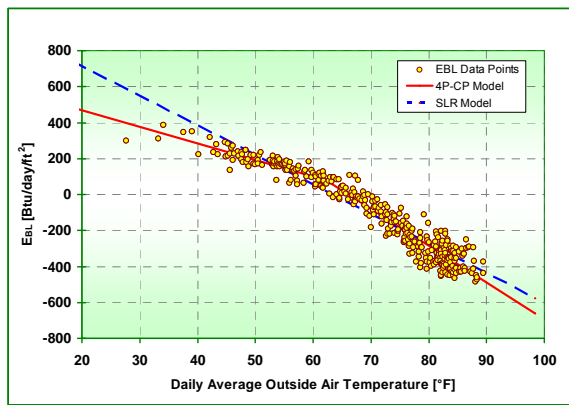


Figure 3 Comparison of 4P-CP and SLR regression models for E_{BL}

METHODOLOGY TO FORMULATE PREDICTION ERROR OF ENERGY BALANCE PARAMETER

As discussed in the previous part, a control chart consists of the mean of samples and the standard deviation. Meanwhile, a regression model estimates the population mean based on samples, and the deviation of the individual prediction by the regression model can be estimated as prediction error. If the definition of control charts is incorporated in regression analysis, the estimated mean response of E_{BL} in Eq. 4 can be interpreted as a regression model. The standard deviation of E_{BL} will be the standard error in predicting new E_{BL} observations corresponding to a specific T by the regression model. For SLR models, the variance of an individual prediction, which is the squared standard prediction error, can be written as Eq. 6 by using symbols relevant to E_{BL} .

$$(S_{\hat{E}_{BL,j}})^2 = \text{MSE} \left[1 + \frac{1}{n} + \frac{(T_j - \bar{T})^2}{\sum_{i=1}^n (T_i - \bar{T})^2} \right] \quad (6a)$$

where

$$\text{MSE} = \frac{\sum_{i=1}^n (E_{BL,i} - \hat{E}_{BL,i})^2}{n - p} \quad (6b)$$

$(S_{\hat{E}_{BL,j}})^2$: estimated variance of individual E_{BL} prediction

\bar{T} : mean value of T

\hat{E}_{BL} : predicted value of E_{BL}

n : number of observations

p : number of parameters in model

i : index for the data during the reference period

j : index for the new observations

However, 4P-CP model is not linear for all range of temperatures, and Eq. 6 does not apply. Figure 4 is the residual plot for the 4P-CP model shown in Figure 3. The residual plot exhibits that the variance is independent of the outside air temperature in the lower temperature region while it increases with the outside air temperature in the higher temperature region. The temperature where the behavior of the residual variance changes is usually close to the change point temperature T_{CP} of the 4P-CP model. In the region higher than this temperature, latent cooling loads become significant and E_{BL} loses linearity. The major reason for this may be the effect of latent load which is not accounted by the model. In the climate such that the

higher temperature is associated with the higher humidity, larger latent load leads to increase in residuals in the higher temperature region because the model used here is a function of outside air dry-bulb temperature. However, Eq. 6 assumes random residuals, independent of outside air temperature. Then the prediction error calculated by Eq. 6 tends to be too large at the lower temperature region and too small at the higher temperature region.

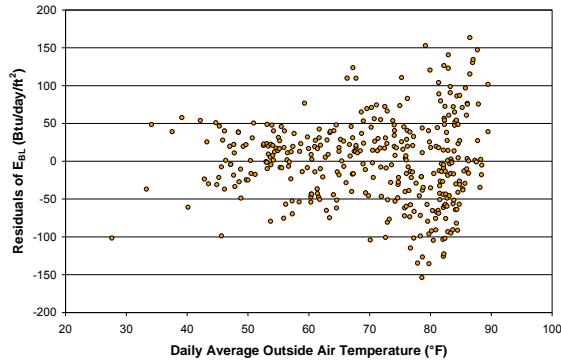


Figure 4 Residual plot for the 4P-CP regression model presented on Figure 3.

Reddy et al. (1997) suggested a simplified approach to change point model with non-uniform residual variance by treating two segments parted by the change point separately, and the residual variance approximated by two variances on either side of the change point has been formulated (Reddy et al. 1998). Let n_1 and n_2 be the number of data points which respectively fall in the lower temperature portion and in the higher temperature portion of the model. Thus:

$$(S_{\hat{E}_{BL,j}})_1^2 = \text{MSE}_1 + \text{MSE} \left[\frac{1}{n} + \frac{(T_j - \bar{T})^2}{\sum_{i=1}^n (T_i - \bar{T})^2} \right] \quad (7a)$$

$$(S_{\hat{E}_{BL,j}})_2^2 = \text{MSE}_2 + \text{MSE} \left[\frac{1}{n} + \frac{(T_j - \bar{T})^2}{\sum_{i=1}^n (T_i - \bar{T})^2} \right] \quad (7b)$$

where

$$\text{MSE}_1 = \frac{\sum_{i=1}^{n_1} (E_{BL,i} - \hat{E}_{BL,i})^2}{n_1} \quad (7c)$$

MSE_2 can be determined from Eq. 7c by analogy. MSE is calculated by Eq. 6b with $p=3$. Although this matches the real case better than Eq. 6a, it may still have the same problem in the higher temperature region.

Control limits constructed by constant or two-level prediction errors results in uneven ability to detect misbehavior in the data along the outside air temperature, which decrease overall performance of the control chart. The new methodology explained next provides a functional expression of prediction error that represents the actual residual model variances well over all range in the outside air temperature.

Residuals variance in E_{BL} (4P-CP) models generally increases as the outside air temperature increases in the higher temperature region. From this behavior, MSE can be assumed as a function of the outside air temperature, and the parameters of the function $\text{MSE}(T)$ may be determined by regression analysis. The procedure to formulate $\text{MSE}(T)$ is described as follows.

- 1) Develop 3°F bin data for E_{BL} and calculate MSE of E_{BL} for each bin, which will be called *Bin-MSE* (BMSE). Let r be the index for a bin and n_r be the number of data in the r^{th} bin. MSE in each bin (BMSE_r) is calculated as:

$$\text{BMSE}_r = \frac{\sum_{i=1}^{n_r} (E_{BL,i} - \hat{E}_{BL,i})^2}{n_r} \quad (8)$$

and the representative temperature of the r^{th} bin, T_r is calculated as:

$$T_r = \frac{1}{n_r} \sum_{i=1}^{n_r} T_i \quad (9)$$

- 2) Split BMSE data into the lower temperature region (BMSE_1) and the higher temperature region (BMSE_2) at the change point temperature, T_{CP} of the E_{BL} model.
- 3) Omit bins that have a small number of data. Then calculate average of BMSE for the lower temperature region (MSE_1) using the following equation.

$$\text{MSE}_1 = \frac{1}{m_1} \sum_{r=1}^{m_1} \text{BMSE}_r \quad (10)$$

where m_1 is the number of BMSE data in the lower temperature region.

- 4) Perform single linear regression on the data set of BMSE_r and T_r in the higher temperature region forcing the line to pass the cross point of MSE_1 and T_{CP} , and find the slope (a) of the line

The resultant function can be expressed in the following equation.

$$MSE(T) = MSE_1 + a(T - T_{CP})^+ \quad (11)$$

Note that if a is zero, in other words, if the $BMSE$ in the higher temperature region does not have a correlation with T , $MSE(T)$ is constant and equal to MSE_1 , which is close to the MSE calculated by Eq. 6b. Figure 5 is the scatter plot of $BMSE$ data with the corresponding $MSE(T)$ model.

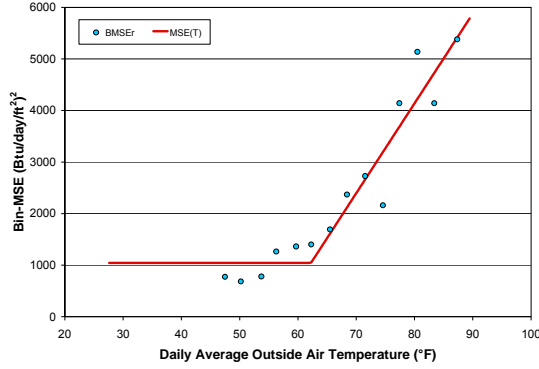


Figure 5 $Bin-MSE$ data points and $MSE(T)$ pattern.

The variance of an individual prediction based on the variable function $MSE(T)$ can be obtained by extending Eq. 7 as:

$$(S_{\hat{E}_{BL,j}}(T_j))^2 = MSE(T_j) + MSE \left[\frac{1}{n} + \frac{(T_j - \bar{T})^2}{\sum_{i=1}^n (T_i - \bar{T})^2} \right] \quad (12a)$$

where

$$MSE = \frac{\sum_{i=1}^n (E_{BL,i} - \hat{E}_{BL,i})^2}{n - p} \quad (12b)$$

n : total number of observations

p : number of parameters in the model

and the standard error in predicting new observations $E_{BL,j}$ at a particular outside air temperature T_j is the square-root of the variance estimated by Eq. 12a. Figure 6 compares residuals of 4P-CP E_{BL} model shown in Figure 4 and 95% confidence prediction intervals estimated by three different ways of calculating variance of individual predictions: 1) constant variance (Eq. 6), 2) 2-level variance (Eq. 7), and 3) variable variance in the higher temperature region (Eq. 12). Although it is not strictly accurate in the statistical sense, the variable variance developed by the proposed methodology fits the actual data which has non-uniform residuals better than those by previous studies especially in the higher temperature region, and it would provide a more effective data screening tool.

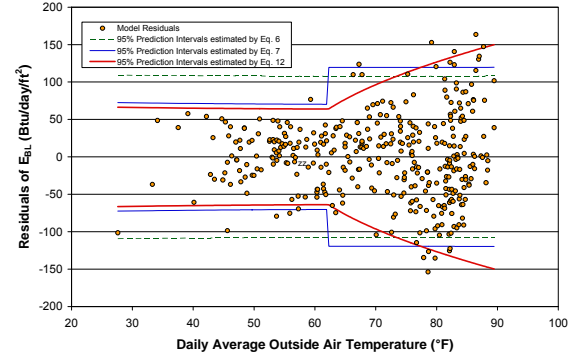


Figure 6 Comparison of 95% prediction intervals calculated by constant variance, 2-level variance, and variable variance plotted over the residual plot

CONSTRUCTION OF CONTROL LIMITS FOR ENERGY BALANCE PARAMETER

The equations to estimate control limits can be obtained in continuous manner by substituting model parameters Eq. 5 and prediction errors Eq. 12 presented in the previous sections into the following equations adapted from Eq. 4, and omitting the subscript j .

$$UCL(T) = \hat{E}_{BL}(T) + kS_{\hat{E}_{BL}}(T) \quad (13a)$$

$$CL(T) = \hat{E}_{BL}(T) \quad (13b)$$

$$LCL(T) = \hat{E}_{BL}(T) - kS_{\hat{E}_{BL}}(T) \quad (13c)$$

UCL and LCL are to be chosen so that practically all the points of the reference data fall within them. Under the assumption that the observations are independent and follow a normal distribution, UCL and LCL are apart from the CL by k times the standard error from CL. If a t -distribution is assumed, k corresponds to t -statistic with the degree of freedom $n-p$. Table 1 compares the fraction of data that are likely to fall within the range defined by the distance of integer multiplier of $\pm S_{E_{BL}}$ from the mean when t distribution is assumed about the mean with the degree of freedom for one year data ($n=362$, $df = n-4$) and the fraction of actual one-year long E_{BL} data ($n=362$) that fall inside the interval calculated by Eq. 13 with $k=1, 2$, and 3 . This shows, in practice, that the coverage factor k is close to t -statistic, and that $k=2$ could be a reasonable number for the control limits by adopting the conventional 95% confidence level. This means that the value of coverage factor represents the level of uncertainty prediction.

Table 1 Fraction of the data falling within the ranges defined by integral multiple of the standard deviation

	$\pm 1S_{EBL}$	$\pm 2S_{EBL}$	$\pm 3S_{EBL}$
1) t-distribution ($df=358$)	$t=1$ 68.2 %	$t=2$ 95.4 %	$t=3$ 99.7 %
2) E_{BL} of actual data ($n=362$)	$k=1$ 71.8 %	$k=2$ 94.2 %	$k=3$ 100 %

Figure 7 is the constructed control limits of E_{BL} plotted in conjunction with the reference E_{BL} data points.

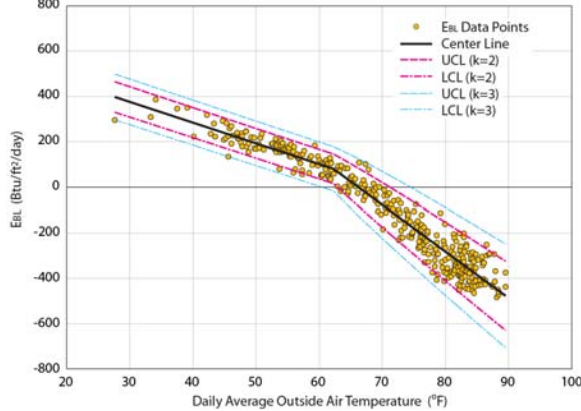


Figure 7 Control Limits of E_{BL}

CASE STUDY

In this section, the algorithms proposed in the previous sections are examined by constructing control limits of E_{BL} for two campus buildings that have different E_{BL} characteristics. Case I presents an application to an offices and laboratory building. This building has 96,038 ft² area and the energy use index (EUI) during the analyzed period was 185 [MBtu/ft²]. The Energy Balance plot shows the variance is dependent of outside air temperature clearly. Case II presents an application to a residence building. The area of this building is 59,541 ft² and the EUI during the analyzed period was 204 [MBtu/ft²]. The Energy Balance plot shows the variance doesn't have a dependency on outside air temperature. The Energy Balance plot for Case I has a steeper line than that for Case II. This indicates Case I building may have larger outside air intake flow rate than Case II building, and this might have caused small E_{BL} variation of Case II building in the higher temperature region. (EUI given here is the annual total of electricity, chilled water, and heating hot water consumptions [MBtu] divided by its total square feet.)

CASE I:

This building is an office and lab building which has a typical data distribution for E_{BL} . The reference data period is 6/1/2005 – 5/31/2006 and the total number of the data points are 364. 4P-CP model equation was found as:

$$E_{BL}(T) = 14.4 + 10.9(66.0-T)^+ - 20.0(T-66.0)^+ \quad [\text{Btu/day-ft}^2] \quad (14)$$

and the $RMSE$ of the model is 44.5. The variance of the data increases as the outside air temperature increases. Then variable MSE as a function of outside air temperature, $MSE(T)$, was estimated by the procedure presented in the previous section. Bin MSE was estimated for 17 temperature bins as given in Table 2. The first three bins were eliminated from the calculation due to the shortage of the number of data in the bins. Figure 8 shows the estimated function $MSE(T)$ and the corresponding $Bin-MSE$ data points. The equation of $MSE(T)$ is:

$$MSE(T) = 390.6 + 190.1(T-66.0)^+ \quad [(\text{Btu/day-ft}^2)^2] \quad (15)$$

The control limits for the case I at the coverage factor $k=2$ and $k=3$ are shown in Figure 9. The reference data points used for developing the control limits were plotted in the same chart.

Table 2 Bin data for the case I

Bin Number r	Temperature Range (°F)	Number of data n_r	Bin T T_r (°F)	Bin MSE $BMSE_r$ (Btu/ft ² /day) ²
1	T<40	5	34.34	357.17
2	40=<T<43	3	41.48	216.80
3	43=<T<46	9	44.54	87.33
4	46=<T<49	15	47.48	404.04
5	49=<T<52	9	50.19	250.56
6	52=<T<55	21	53.73	291.94
7	55=<T<58	20	56.28	332.41
8	58=<T<61	15	59.69	473.56
9	61=<T<64	19	62.27	591.22
10	64=<T<67	18	65.50	873.09
11	67=<T<70	22	68.42	1438.63
12	70=<T<73	25	71.52	2374.76
13	73=<T<76	24	74.58	2845.13
14	76=<T<79	30	77.40	3671.52
15	79=<T<82	23	80.69	4257.24
16	82=<T<85	74	83.44	2341.69
17	85=<T<88	32	87.25	3602.93

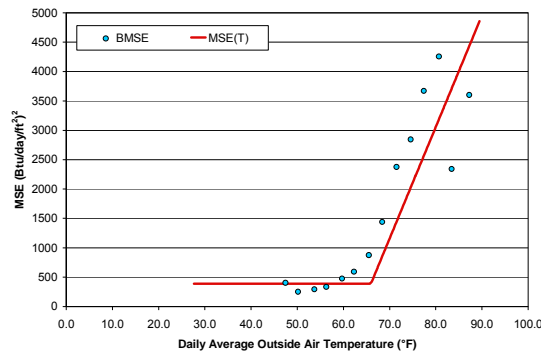


Figure 8 Bin-MSE data points and $MSE(T)$ for the case I

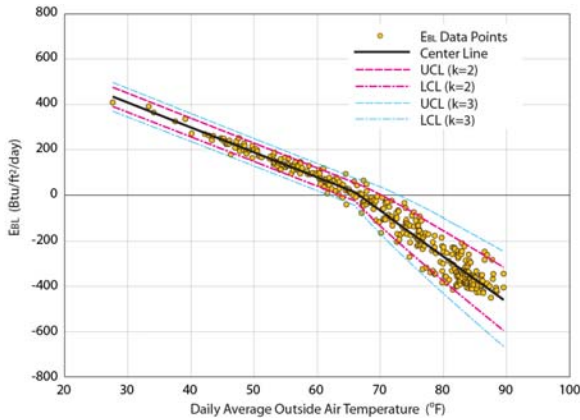


Figure 9 Control Limits of E_{BL} for the case I

CASE II:

This building is a residence building which has rather constant variance of E_{BL} . The reference data period is 11/1/2006 – 10/31/2007 and the total number of the data points are 364. 4P-CP model equation was found as:

$$E_{BL}(T) = -24.5 + 5.7(65.8-T)^+ - 8.1(T-65.8)^+ \quad [\text{Btu/day-ft}^2] \quad (16)$$

and the $RMSE$ of the model is 21.3. In this case, the variance in the E_{BL} does not seem to have a correlation with the outside air temperature. Bin-MSE data was estimated as in Table 3. It shows scattered $BMSE$ values and no correlation with outside air temperature. Then instead of using variable MSE , the classic MSE which assumes constant variance was used, and the prediction error was estimated by Eq. 6. Figure 10 shows Bin-MSE data points and $MSE(T)$ which is constant and equal to MSE estimated using all the data. The control limits at the coverage factor $k=2$ and $k=3$ are shown in Figure 11.

Table 3 Bin data for the case II

Bin Number r	Temperature Range (°F)	Number of data n_r	Bin T T_r (°F)	Bin MSE $BMSE_r$ (Btu/ft²/day)²
1	$T < 40$	11	35.51	383.12
2	$40 \leq T < 43$	10	41.42	652.03
3	$43 \leq T < 46$	14	44.44	771.69
4	$46 \leq T < 49$	14	47.73	338.62
5	$49 \leq T < 52$	14	50.54	579.38
6	$52 \leq T < 55$	12	53.54	563.06
7	$55 \leq T < 58$	15	56.17	311.33
8	$58 \leq T < 61$	18	59.52	592.16
9	$61 \leq T < 64$	17	62.48	615.81
10	$64 \leq T < 67$	23	65.56	822.35
11	$67 \leq T < 70$	17	68.69	505.43
12	$70 \leq T < 73$	34	71.48	502.21
13	$73 \leq T < 76$	20	74.48	827.34
14	$76 \leq T < 79$	35	77.50	365.27
15	$79 \leq T < 82$	58	80.46	216.60
16	$82 \leq T < 85$	42	83.47	158.63
17	$85 \leq T < 88$	10	87.44	426.87

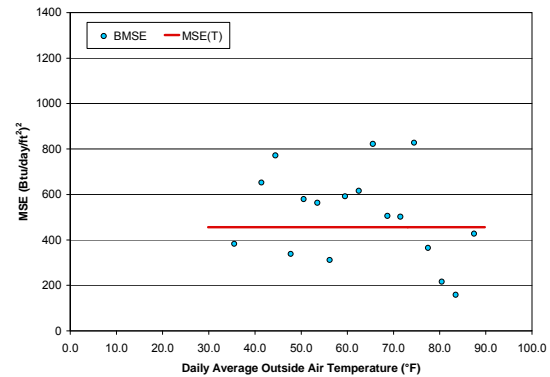


Figure 10 Bin-MSE data points and $MSE(T)$ for the case II

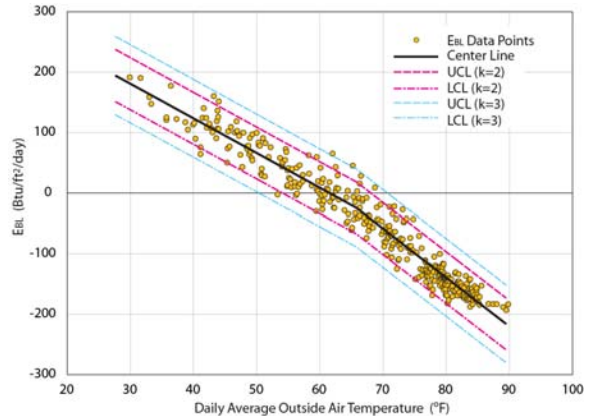


Figure 11 Control Limits of E_{BL} for the case II

CONCLUSIONS

This paper presented the methodology to develop control limits for data quality control using Energy Balance methodology. The new method was introduced to solve the problem that the Energy

Balance parameter has non-uniform model residuals, by expressing prediction error as a function of outside air temperature with practical approach. The numerical criteria for data screening of E_{BL} determined by this method would provide more uniform and stable results in data quality control than depending on visual inspection. In addition, the level of the control limits can be indicated associated with the uncertainty level, which standardizes the interpretation of the results in the data screening process.

To determine reasonable control limits by this method, the whole building level of energy use data have to be measured separately for electricity, chilled water, and heating hot water, and the data period should be at least one year. When the operation condition for the building is changed, new control limits should be developed using the data collected under the new condition. The bin size selected in this paper was 3 °F and the number of *Bin-MSE* data used for developing variable MSE model was 14 in Case I and 17 in Case II, respectively. The larger bin size such as 5°F bin didn't change the result remarkably. However, if 5 °F bin is used, the number of the *Bin-MSE* data is reduced to around 10, and this may be a disadvantage in using regression analysis.

It was shown that, in a case E_{BL} has non-uniform model residuals, this method have better results in estimating prediction errors that has uniform fitting overall temperature range than those in previous studies. This provides more effective control limits for data screening. However, Bin MSE data is sensitive to large or small values, especially when the number of the data in the bin is not large enough. This would effect on resulting MSE functions significantly. Further case study and statistical discussion will be needed to increase the robustness in this methodology.

REFERENCES

- Baltazar, J. C., Sakurai, Y., Masuda, H., Feinauer, D., Liu, J., Ji, J., Claridge, D. E., Deng, S. and Bruner, H., (2007), "Experiences on the Implementation of the "Energy Balance Methodology" Methodology as a Data Quality Control Tool: Application to the Building Energy Consumption of a Large University Campus", *International Conference for Enhanced Building Operations*, San Francisco, CA, Nov.
- Baltazar, J.C., (1999), *CPReg: Change Point Regression Tool*, Energy Systems Laboratory, College Station, TX.
- Kissock, J. K., Haberl, J.S., and Claridge, D.E., (2002), Development of a Toolkit for Calculating Linear, Change-point Linear and Multiple-Linear Inverse Building Energy Analysis Models. *ASHRAE Research Project 1050-RP: Final Report*. Nov.
- Kissock, J. K., (2007), EModel 1.4F: a Microsoft VBA tool for calculating building energy use model, Energy Systems Laboratory, College Station, TX.
- Montgomery, D.C. and Runger, G.C., (2003), *Applied Statistics and Probability for Engineers*, 3rd Ed. John Wiley & Sons, New York, NY.
- Reddy, T.A., Kissock, J.K., and Ruch, D.K., (1998), "Uncertainty in Baseline Regression Modeling and in Determination of Retrofit Savings", *Journal of Solar Energy Engineering*. Vol. 120. pp. 185-190. Aug.
- Reddy, T.A., Saman, N.F., Claridge, D.E., Haberl, J.S., Turner, W.D., and Chalifoux, A.T., (1997), "Baselining Methodology for Facility-Level Monthly Energy Use – Part 1, Theoretical Aspects", *ASHRAE Transactions*, Vol. 103. pt. 2. pp.336-347.
- Ruch, D. and Claridge, D.E., (1992), "A Four-Parameter Change-Point Model for Predicting Energy Consumption in Commercial Buildings", *Journal of Solar Energy Engineering*, Vol. 114. pp.77-83, May.
- Shao, X. and Claridge, D.E. (2006), "Use of First Law Energy Balance as a Screening Tool for Building Energy Data: Part I – Methodology", *ASHRAE Transactions*. Vol. 112. Part 2.
- Shao, X. (2005). First Law Energy Balance as a Data Screening Tool, MS Thesis, Mechanical Engineering Department, Texas A&M University. College Station, TX, May.